Lower dimensional kernels for video discriminators

Neural Networks 132 (2020) 506–520

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2020 Special Issue

Lower dimensional kernels for video discriminators☆

Emmanuel Kahembwe [a,b,c,*], Subramanian Ramamoorthy [a,b,d]

[a] Robust Autonomy and Decisions Group, The School of Informatics, The University of Edinburgh, 10 Crichton St, Edinburgh EH8 9AB, United Kingdom
[b] The Edinburgh Centre of Robotics, The University of Edinburgh's Bayes Centre, 47 Potterrow, Edinburgh EH8 9BT, United Kingdom
[c] The School of Engineering and Physical Sciences, The Robotarium, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom
[d] FiveAI, 5th Floor, Greenside, 12 Blenheim Place, Edinburgh, EH7 5JH, United Kingdom

ARTICLE INFO

Article history:
Available online 26 September 2020

Keywords:
Generative Adversarial Networks
Discriminator analysis
Video generation

ABSTRACT

This work presents an analysis of the discriminators used in Generative Adversarial Networks (GANs) for Video. We show that unconstrained video discriminator architectures induce a loss surface with high curvature which make optimization difficult. We also show that this curvature becomes more extreme as the maximal kernel dimension of video discriminators increases. With these observations in hand, we propose a methodology for the design of a family of efficient Lower-Dimensional Video Discriminators for GANs (LDVD-GANs). The proposed methodology improves the performance and efficiency of video GAN models it is applied to and demonstrates good performance on complex and diverse datasets such as UCF-101. In particular, we show that LDVDs can double the performance of Temporal-GANs and provide for state-of-the-art performance on a single GPU using the proposed methodology.

1. Introduction

1.1. Motivation

Most high-dimensional datasets of interest, involving images and video, lie on some lower dimensional manifold. Density estimation of these datasets presents a unique challenge that has very few practical solutions. Generative adversarial networks (GANs) are an implicit density estimation approach to unsupervised learning that provide for one such solution. Schmidhuber (1990, 1991) laid out the fundamental principles underpinning adversarial training of neural networks, as elaborated in Schmidhuber (2020). Goodfellow et al. (2014) introduced the GAN framework that extended these principles to the modelling of complex high-dimensional data distributions. In the case of images, GANs provide a mapping, $G : z \rightarrow X$, from a lower dimensional latent code, $z \in \mathbb{R}^d$, to the high dimensional image manifold, $X \in \mathbb{R}^{h \times w}$. During training, they use an auxiliary network, D, to minimize the Jensen–Shannon divergence between the distribution of images induced by the mapping G and the true data distribution of real images from the dataset. The mapping G uses the gradient signal from D to improve the quality of images it produces until they are indistinguishable from the real images from the true data distribution.

The efficacy of this training regime has revolutionized the field of image generation, subsequently establishing GANs as a leading method for image generation. GAN-based image generation has observed significant advances; from architectural contributions (Karras, Aila, Laine, & Lehtinen, 2018; Radford, Metz, & Chintala, 2016; Zhang, Goodfellow, Metaxas, & Odena, 2019) to novel forms of losses (Arjovsky, Chintala, & Bottou, 2017; Jolicoeur-Martineau, 2019) and stabilization methods (Gulrajani, Ahmed, Arjovsky, Dumoulin, & Courville, 2017; Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017; Miyato, Kataoka, Koyama, & Yoshida, 2018; Salimans et al., 2016). Current state-of-the-art models for image generation produce high-resolution visual results that are sometimes difficult for humans to distinguish from those derived from the true data distribution (Brock, Donahue, & Simonyan, 2019; Karras, Laine, & Aila, 2019).

In comparison, video generation has not enjoyed the same level of progress as image generation. Video GAN (VGAN) models have benefited from methods such as progressive growing (Acharya, Huang, Paudel, & Gool, 2018) and Lipschitz regulation (Saito, Matsumoto, & Saito, 2017). However, VGAN models are still insufficient when it comes to modelling the true video distribution and produce results that are easily identified as lying outside its support.

☆ This research is supported by the Engineering and Physical Sciences Research Council (EPSRC), United Kingdom, as part of the CDT in Robotics and Autonomous Systems at Heriot-Watt University and The University of Edinburgh, Grant reference EP/L016834/1.

* Corresponding author at: Robust Autonomy and Decisions Group, The School of Informatics, The University of Edinburgh, 10 Crichton St, Edinburgh EH8 9AB, United Kingdom.
E-mail address: e.kahembwe@ed.ac.uk (E. Kahembwe).

https://doi.org/10.1016/j.neunet.2020.09.016

There are many possible reasons for the comparatively limited performance of GAN models on the task of video modelling when compared to image modelling. The most obvious reason being that video modelling is a higher dimensional and more complex task due to the addition of a temporal dimension. The additional dimension significantly increases the number of parameters required by a GAN model to sufficiently capture the true data distribution, resulting in higher memory and compute costs.

1.2. Summary

We analyse the discriminators used in VGAN models and reveal a more nuanced contributing factor to the limited performance of these models. We find that the dimensionality of the 3D kernels used in video discriminators induces significantly higher curvature in the loss landscape when compared to that induced by 2D kernels and that this is detrimental to first order optimization methods such as stochastic gradient decent. Although gradient decent is known to require a smooth loss landscape for good performance, the link between kernel dimensionality and the curvature of the loss landscape was unknown prior to this work. We discover and explore this link, which results in an explanation for a series of anomalous behaviours and characteristics exhibited by video GANs. In light of this observation, we target the maximum kernel dimensionality of video GAN models as an area of optimization for the purposes of stabilizing training and improving performance. We develop a video discriminator design methodology which can be used to improve the performance of any video GAN model and demonstrate our methodology on two existing video GAN models. We also explore computation and memory efficiency from this perspective. To conclude, we demonstrate that in lowering the maximal kernel dimensionality of a video GAN model, we can reduce pathologies in the loss landscape, improve overall model performance while significantly increasing memory and computation efficiency. The proposed design principles improve the performance of every video GAN model we apply them to and provide for state-of-the-art results at a level of efficiency unmatched by any existing video GAN model.

Our novel contributions are as follows:

- We provide the first analysis of video discriminators.
- We discover a previously unknown phenomenon; that the curvature of the loss landscape for video GANs deteriorates as the convolution kernel dimensionality of their discriminator increases.
- We propose a set of discriminator design principles, that when applied together lead to improved performance and efficiency of video GAN models.
- Application of our proposed methodology to common video GAN models provides for state-of-the-art single-gpu results on the UCF-101 dataset.
- We show the first video GAN model capable of modelling video at resolutions of up to 512×512.

1.3. Layout

This work is organized as follows; in the following section we review the relevant literature in both the image and video GAN domains. Section 3 details an analysis of the MoCoGAN and TGAN video discriminators. Section 4 introduces a family of lower dimensional video discriminators. Section 5 compares the performance of the different discriminators against prior art. Section 6 concludes this study with a discussion of the results and its implications for the design of video discriminators.

2. Related work

2.1. Image generation

Adversarial training, within the context of neural networks (Goodfellow et al., 2014; Schmidhuber, 1990, 1991, 2020), pits two networks against each other in a zero-sum non-cooperative game which is solved, in the game-theoretic sense, by the application of the minimax theorem (v. Neumann, 1928). A network, called the generator (G), learns to model the true data distribution, p_{data}, by fooling an adversary, termed the discriminator (D), whose job is to learn a classifier that tells apart the generated data from the real data. The standard formulation of this game is given by the value function $V(G, D)$;

$$\min_{\theta} \max_{\psi} V(G, D) = \mathbb{E}_{x \sim p_{data}(x)}\left[f_D\left(D_\psi(x)\right)\right] + \mathbb{E}_{z \sim p_z(z)}\left[f_G\left(D_\psi\left(G_\theta(z)\right)\right)\right] \tag{1}$$

where ψ and θ are parameters for D and G respectively, p_z is a prior distribution and $f(\cdot)$ is a real-valued function whose exact form depends on the choice of loss function (Arjovsky et al., 2017; Goodfellow et al., 2014; Nowozin, Cseke, & Tomioka, 2016). Goodfellow et al. (2014) initially presented models trained with this approach, termed Generative Adversarial Networks (GANs), on the task of image generation. In the GAN research community, significant effort has been dedicated to improving performance on image generation tasks as a benchmark and to stabilizing the adversarial training regime (Arjovsky et al., 2017; Brock et al., 2019; Miyato et al., 2018; Radford et al., 2016; Zhang et al., 2019). To facilitate evaluation of GAN research, metrics such as the Inception Score (IS) (Salimans et al., 2016) and the Fréchet Inception Distance (FID) (Heusel et al., 2017) have become the standard for benchmarking image quality and diversity respectively.

2.2. Video generation

Generative adversarial networks for video (VGAN) by Vondrick, Pirsiavash, and Torralba (2016) was the first model to extend GANs to the video domain. It is composed of a 3D discriminator and a two-stream generator; one stream generating the static background content using 2D kernels and another 3D stream generating the foreground dynamic content.

Subsequent models such as Temporal GAN (TGAN) (Saito et al., 2017) and Motion and Content decomposed GAN (MoCoGAN) (Tulyakov, Liu, Yang, & Kautz, 2018) use a two-stage generator. In TGAN, the first stage samples a random latent $z_c \in \mathbb{R}^{d_c}$ and generates a set of latent variables $z_m^{1..T}$ conditioned on z_c that define a video trajectory across T time-steps. The second stage takes the concatenated latents $[z_c; z_m^t]$,[1] and generates a single image for each time-step $t \in T$. The first stage of TGAN's generator uses 1D kernels, followed by 2D kernels for image generation in the second stage. It utilizes 3D kernels in its discriminator.

MoCoGAN (Tulyakov et al., 2018) explicitly models the static and dynamic attributes of video separately. This model assumes a factorized latent space, where the content in video is embedded on a subspace, $z_c \in \mathbb{R}^{d_c}$ and its associated motion is embedded on another subspace, $z_m \in \mathbb{R}^{d_m}$. The generator models the joint space $Z_l \in \mathbb{R}^d$, where $\mathbb{R}^d = \mathbb{R}^{d_m} + \mathbb{R}^{d_c}$, using a combination of a Gated Recurrent Unit (GRU) (Cho et al., 2014) to generate the temporal dynamics of video and a 2D convolutional upscaling network to generate the associated spatial information. The video generation process involves sampling a latent variable z_m^t at each time-step $t \in T$ from the motion subspace, processing it with a

[1] $[z_c; z_m]$ denotes the concatenation operation; between z_c and z_m.

GRU and concatenating the resulting output with z_c to form $z = [z_c; GRU(z_m^{1-T})]$, where $z \in Z_l$. The upscaling network is then used to project z to the image space to form video frames. The MoCoGAN discriminator architecture consists of two components, a 2D image and 3D video discriminator. It is the current state-of-the-art model for low-resolution, 64 × 64 video generation (Soomro, Zamir, & Shah, 2012).[2]

In the high-resolution video GAN literature, Acharya et al. (2018) propose Progressive Video GAN (ProVGAN), a model that combines progressive growing with 3D kernels and a sliced-Wasserstein loss to generate video at 256 × 256 resolution. Saito and Saito (2018) explore sub-sampling as a way to scale video GAN models to 192 × 192 video generation and achieve state-of-the-art results when combined with a large batch training regime. Their proposed model, TGANv2,[3] reduces memory and computational costs by sub-sampling across time, space and batch size as the resolution of feature maps increases. The TGANv2 video generator is comprised of a convolutional LSTM and a 2D image generator. It also uses multiple rendering layers at different resolutions within the generation pipeline, in conjunction with a hierarchy of 3D discriminators to critic the generated video at different spatial and temporal resolutions. It is the state-of-the-art model for high resolution video generation.

3. Discriminator architectures

Although there is some variety in the architectures of video GAN generators, the associated discriminator architectures have remained fairly consistent. There are currently two primary choices for video GAN discriminators; a dual (2D image + 3D video) discriminator architecture as in MoCoGAN, or a 3D video discriminator architecture as in VGAN, TGAN and ProVGAN.[4] Although it is relatively easy to develop well performing video GAN generators, changes to video discriminators often lead to GANs that perform badly and in many cases, cannot even train. This has culminated in video discriminators being the most computational and memory expensive components in video GAN models.

In this section, we take a closer look at the current choice of video GAN discriminators and study the impact that architectural decisions for this component have on model performance. In particular, we explore the question;

- *"What makes a good discriminator for video GANs?"*.

To answer this question, we analyse the properties of seminal video GAN discriminators for both dual and single discriminator architectures. For the dual discriminator architecture, our analysis focuses on the MoCoGAN model since it is the first video generation model to incorporate multiple architectural components in its discriminator. TGAN is used as the representative model for single component discriminators due to the architectural similarity of its video discriminator to that of MoCoGAN.

3.1. Experimental setup

The original experimental code for the MoCoGAN[5] and TGAN[6] model is publicly available. As such, all experiments use the original experimental code and settings to allow for accurate analysis and aid reproducibility. Training and performance benchmarking is carried out on a single 12GB Titan-X GPU.

3.1.1. Quantitative experiments

We use the UCF-101 dataset for our quantitative experiments (Soomro et al., 2012). This dataset was initially introduced for action recognition but was co-opted by the GAN research field in order to benchmark video GAN models. The UCF-101 dataset (Soomro et al., 2012) is a video dataset consisting of 13,320 video clips divided across 101 different action categories, at a spatial resolution of 320 by 240 pixels.

Our preprocessing pipeline centre crops all videos to 240 × 240 pixels. For the MoCoGAN experiments, the video is temporally subsampled by a factor of 2 and 16 consecutive frame are randomly extracted. For the TGAN experiments, the video is not subsampled in order to match the original experimental conditions.[7] We then resize the resulting video to the model resolution (i.e. 16 × 64 × 64 or 16 × 128 × 128). We use the *"trainlist01"* training split containing 9537 videos to train our models as in previous works.

Evaluation metrics. We benchmark performance against the video extensions of the Inception Score (IS) and Fréchet Inception Distance (FID) where a higher IS implies that the generated video is of a higher visual quality and a lower FID implies a better fit to the modes of the data distribution (i.e. good diversity).[8] These metrics are highly sensitive to implementation so we use their exact implementation from the original TGAN experiments (Saito et al., 2017). We calculate the IS using the Sport1M pre-trained C3D classification model fine-tuned on UCF-101 from the original TGAN experiments. We calculate the FID using the activations from the second to last linear layer, denoted *fc7*, of the same C3D model. As in previous works, we generate 10000 samples from the model to evaluate each metric and derive a rough standard deviation by repeating this procedure four times.

3.1.2. Qualitative results

We use the MUG Facial Expression Database (MUG-FED) for our qualitative experiments (Aifanti, Papachristou, & Delopoulos, 2010). This dataset is composed of 1462 video sequences of 86 people demonstrating 7 categories of facial expression; anger, fear, disgust, happiness, sadness, surprise and neutral. The videos are recorded at a resolution of 896 × 896 pixels and range between 40–180 frames.

Evaluation metrics. For each model, 10000 video samples are generated and randomly sub-divided into 100 batches. All videos in each batch are tiled and aggregated into a single larger video. The tiled videos are presented, two at a time, to 12 human participants for a side-by-side visual evaluation of sample quality and batch diversity.

3.2. MoCoGAN discriminator

Although the seminal video GAN models used a single 3D video discriminator, later works such as MoCoGAN achieve better results with a dual (2D image + 3D video) discriminator model. In theory, a high-capacity discriminator with kernels whose dimensionality matches that of the input data distribution should allow for better criticism. In practice, this is not observed and the MoCoGAN authors attribute the performance boost gained with the addition of an image level discriminator to its ability to "focus

[2] We refer to the state-of-the-art as benchmarked on the UCF-101 dataset.

[3] TGANv2 (Saito & Saito, 2018) was published while this work was in review, we have included it in this article post hoc.

[4] The dimensionality of the discriminator is always in reference to the maximum convolution kernel dimension.

[5] MoCoGAN Code: https://github.com/sergeytulyakov/mocogan.

[6] TGAN Code: https://github.com/pfnet-research/tgan.

[7] Results for TGAN trained with temporal subsampling and MoCoGAN trained without temporal subsampling are explicitly labelled.

[8] These metrics are defined relatively and should be approached with caution (Barratt & Sharma, 2018).

Table 1
MoCoGAN Image and Video Discriminators.

Layer	Image Configuration	Video Configuration
Input	height × width × 3	16 × height × width × 3
c0	Conv2D-(N64, K4, S2, P1), LReLU	Conv3D-(N64, K4, S(1,2,2), P(0,1,1)), LReLU
c1	Conv2D-(N128, K4, S2, P1), BN, LReLU	Conv3D-(N128, K4, S(1,2,2), P(0,1,1)), BN, LReLU
c2	Conv2D-(N256, K4, S2, P1), BN, LReLU	Conv3D-(N256, K4, S(1,2,2), P(0,1,1)), BN, LReLU
c3	Conv2D-(N1, K4, S2, P1)	Conv3D-(N1, K4, S(1,2,2), P(0,1,1))

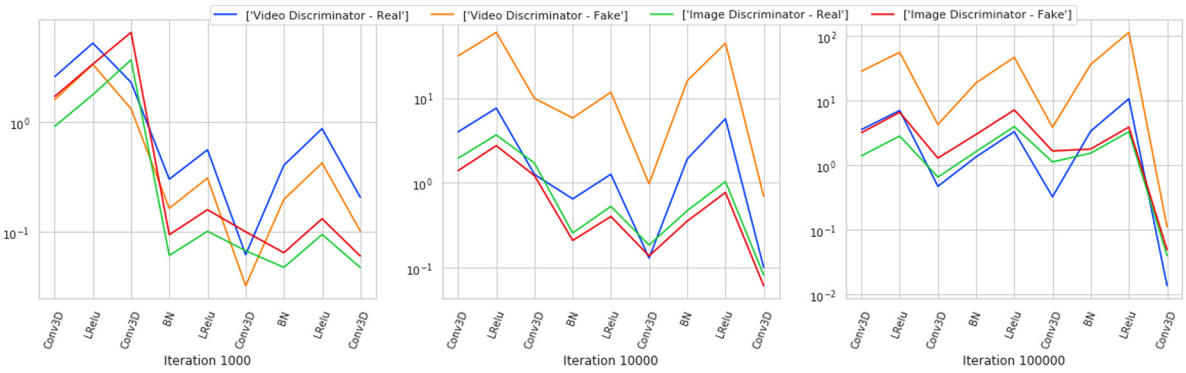


Fig. 1. Norm of the gradient at each node in the computation graph for batches of real and fake data.

on static appearances". In the following section, we investigate this claim and provide a more thorough explanation grounded with empirical observations.

3.2.1. Ablation study

Table 1 details the architectures for the MoCoGAN 2D image and 3D video discriminators. These are identical patch-level discriminators that operate on 46 × 46 patches of the input video frames with the distinction that the image discriminator is comprised of 2D convolution kernels and the video discriminator utilizes 3D kernels.

We replicate the results for this model on the UCF-101 dataset and ablate its discriminator components to ascertain how much each component contributes to the final model performance. Results are presented in Table 2.

In Table 2, we observe that there is a degradation in performance without an image-level discriminator (row 1 vs row 3) and we also observe that image-level statistics can account for most of the model performance as demonstrated by the model trained with an image-only discriminator (row 4).

We were not able to replicate the published results and suspect that this could be due to the MoCoGANs saturating loss function. The loss function used in the original GAN formulation set $f_D(t) = \log(t)$ and $f_G(t) = \log(1-t)$ in Eq. (1). f_G saturates when the discriminator overpowers the generator and learns a perfect classifier between the real and fake data distribution. As a result, the gradient signal propagated back to the generator via the discriminator vanishes, stalling training for the generator. Goodfellow et al. (2014) proposed a solution for this by training the generator to maximize f_D, i.e. $f_G = -f_D$. This provides for a non-saturating version of the GAN loss that allows for a non-vanishing gradient signal through out training (Fig. 1). The non-saturating loss did not improve performance but we maintain it for all further experiments to avoid potential saturation issues. We were only able to replicate the published performance of the MoCoGAN model, by doubling the number of channels in every component of the model (row 5). Doubling just the channels for the video discriminator was not sufficient even though it accounts for 80% of the discriminator parameters (row 6). Finally, we also measure the impact of temporal subsampling on model performance as it was not used in previous models such as TGAN (row 2). We observe that temporal subsampling accounts for a significant portion of model performance.

3.2.2. Hessian analysis

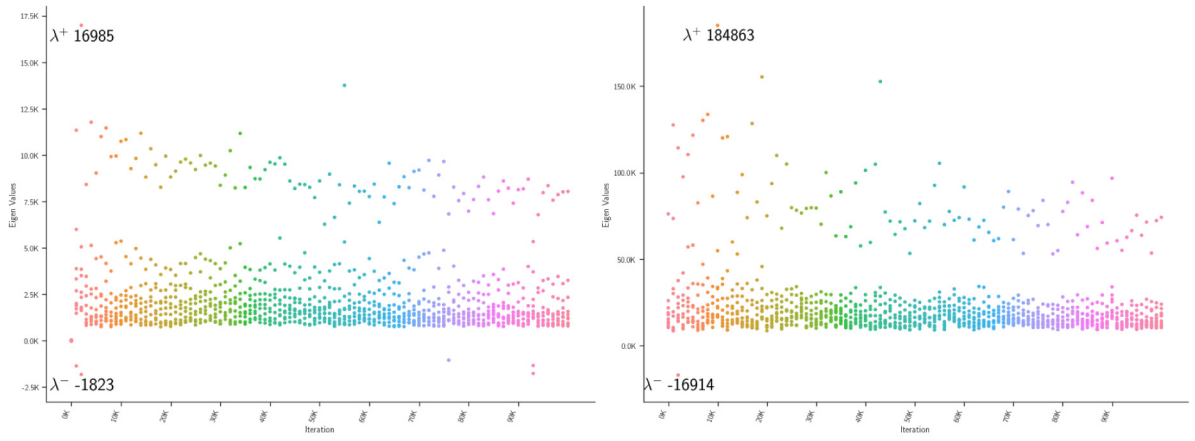
In this section, we analyse the loss surface induced by the image and video discriminators via analysis of the Hessian of the GAN objective with respect to the discriminator parameters. Hessian analysis of neural networks is computationally expensive, especially for large models such as those used in the video GAN domain. As an approximation, we employ the R-operator from Pearlmutter (1994) to calculate the exact Hessian vector product for neural networks and combine it with the Lanczos algorithm to calculate the eigen spectra of the Hessian. The gradient and Hessian are taken with respect to the parameters of the discriminator and are given by $\dot{\psi} = \nabla V(G, D)$ and $\dot{\psi} = \nabla^2 V(G, D)$ respectively. We track the leading eigenvalues of the Hessian throughout training and present these results in Fig. 2.

Fig. 2 provides an interesting perspective as to what is going on with the MoCoGAN discriminator. The primary observation is that for identical discriminator architectures, an increase in kernel dimensionality results in a loss surface with significantly more pathological curvature. Each discriminator component individually encounters eigenvalues during training that are an order of magnitude larger than the majority of leading eigenvalues. But there is also a further order of magnitude difference between the maximum eigenvalue and majority leading eigenvalues for the 2D image discriminator, when compared against the 3D video discriminator (see Figs. 2(a) vs 2(b) and Figs. 2(c) vs 2(d)). An increase in parameters results in a generally smoother loss landscape with possibly more saddle points (see Figs. 2(a) vs 2(c) and Figs. 2(b) vs 2(d)). A consistent observation is that the magnitude of the largest eigenvalue tends to reduce throughout training and that this effect is less observable with an increase in discriminator kernel dimensionality.

Kernel complexity vs kernel dimensionality. It could be said that the observations in Fig. 2 are possibly an artefact of kernel complexity rather than dimensionality. But increasing the parameter complexity of the image discriminator such that it matches that of the video discriminator leads to little deterioration in the loss landscape of the image discriminator (Figs. 3(a) vs 2(b)). Instead

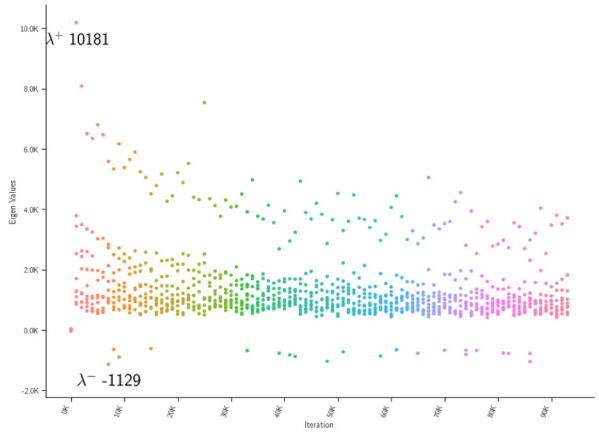
Table 2
MoCoGAN Ablation.

Row	Model	Disc Params	IS ↑	FID ↓
1	MoCoGAN	3.3M	11.58 ± .04	9485.34 ± 14.61
2	MoCoGAN — No Temporal Subsampling	3.3M	10.35 ± .06	9657.44 ± 3.90
3	MoCoGAN — Video Discriminator Only	2.7M	11.09 ± .03	9565.65 ± 21.03
4	MoCoGAN — Image Discriminator Only	.7M	8.26 ± .04	10494.09 ± 12.52
5	MoCoGAN — 2xChannels	13.2M	12.61 ± .08	9166.81 ± 9.92
6	MoCoGAN – 2xChannels – Video Discriminator Only	10.5M	11.73 ± .05	9461.10 ± 1.85
7	MoCoGAN – Video Discriminator Only – ksize2	.3M	3.98 ± .02	13664.81 ± 7.64
8	MoCoGAN – Image Discriminator Only – ksize8	2.7M	7.62 ± .04	10337.71 ± 0.47
9	MoCoGAN — Original published in Tulyakov et al. (2018)	3.3M	12.42 ± .03	

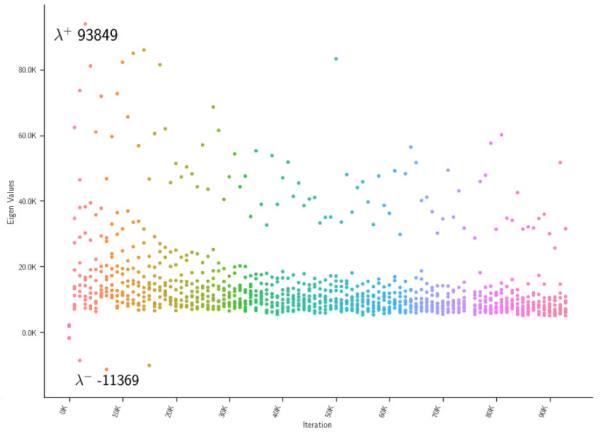


(a) MoCoGAN- Image Discriminator

(b) MoCoGAN- Video Discriminator



(c) MoCoGAN 2xChannels - Image Discriminator

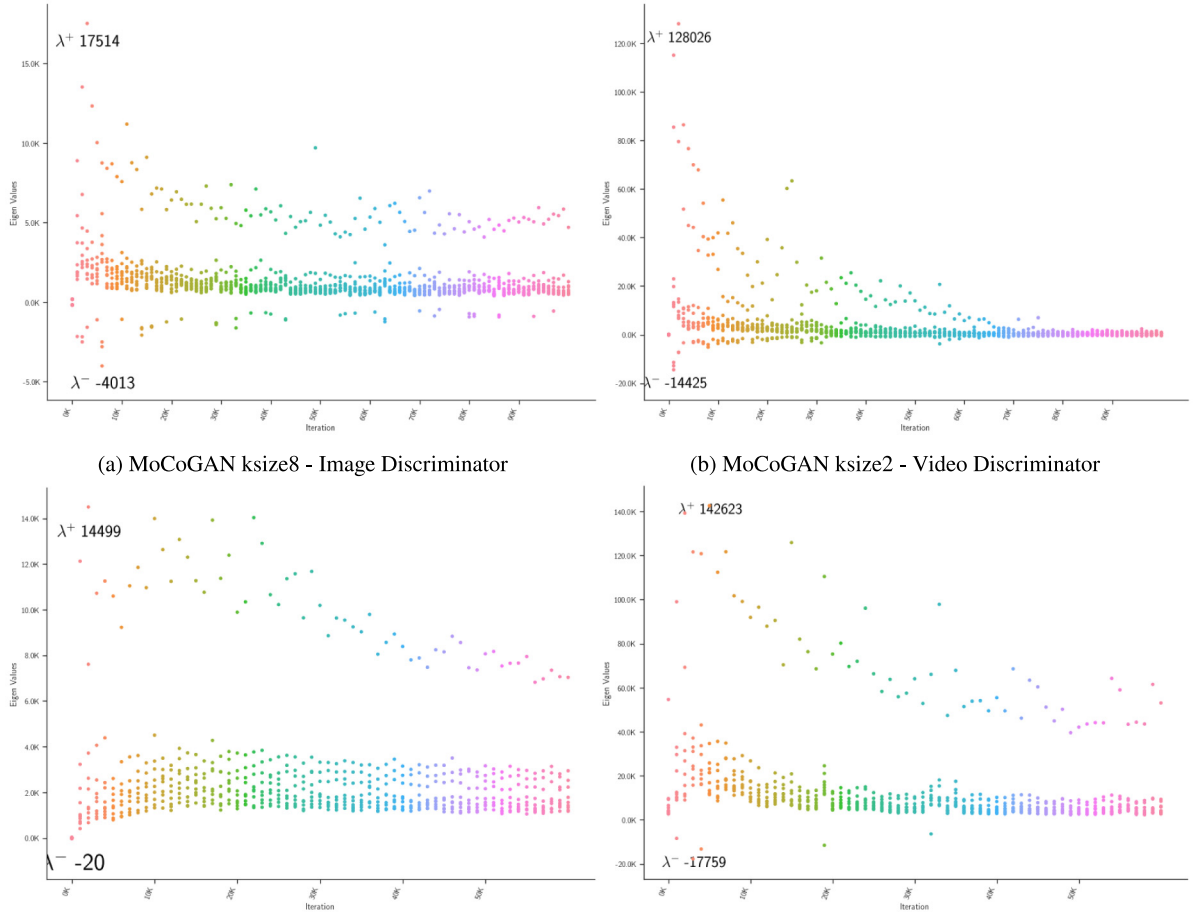


(d) MoCoGAN 2xChannels - Video Discriminator

Fig. 2. The 10 largest eigenvalues of the loss Hessian with respect to the image and video discriminators of the MoCoGAN architecture. (a) and (b) show these values for the original MoCoGAN architecture. (c) and (d) show these for MoCoGAN with the number of channels doubled (see MocoGAN — 2xChannels in Table 2). λ^+ denotes the largest positive eigenvalue encountered during training and λ^- the largest negative eigenvalue.

we observe that even with an increase in kernel complexity, the loss landscapes induced by the image discriminators are more similar to each other in terms of curvature than to those induced by the video discriminator (Figs. 2(a), 2(c), and 3(a) vs 2(b)). Furthermore, Fig. 3(b) shows that before collapse, a video discriminator with half the kernel complexity of an image discriminator induces a similar loss landscape to that of the original MoCoGAN video discriminator (Figs. 2(b) vs 3(b)) .

Dataset complexity vs kernel dimensionality. An argument could be made that the dataset complexity affects the curvature of the loss landscape. We analyse the loss surface of the MoCoGAN model trained on two different datasets with very different mode characteristics and scene dynamics, UCF-101 and MUG-FED. UCF-101 includes 101 different classes with different inter and intra class variability for each video when compared to MUG-FED. MUG-FED is a dataset with a fixed background and different faces under controlled lighting making different facial expressions. UCF-101 is a significantly more complex dataset than MUG-FED and our analysis shows that indeed, the loss landscape is affected by the dataset, resulting in landscapes with different characteristics (see Figs. 2(a) vs 3(c) and Figs. 2(b) vs 3(d)). But we still observe that the curvature of this landscape is primarily dictated by the kernel dimensionality resulting in similar curvature magnitude profiles for models trained on either dataset



(a) MoCoGAN ksize8 - Image Discriminator

(b) MoCoGAN ksize2 - Video Discriminator

(c) MoCoGAN MUG - Image Discriminator

(d) MoCoGAN MUG - Video Discriminator

Fig. 3. The 10 largest eigenvalues of the loss Hessian with respect to the image and video discriminators of the MoCoGAN architecture. (a) shows these values for an image discriminator with kernels that have four times the parameters of the original image discriminator. (b) shows these values for a video discriminator with an eighth the parameters per kernel of the original video discriminator. (c) and (d) show these values for the MoCoGAN discriminators trained on the MUG-FED dataset. λ^+ denotes the largest positive eigenvalue encountered during training and λ^- the largest negative eigenvalue.

(see Figs. 2(a), 2(b) vs Figs. 3(c), 3(d)). In particular, we observe the same order of magnitude difference between the eigenvalues of the loss Hessian for 2D image discriminators when compared against their 3D video counterparts.

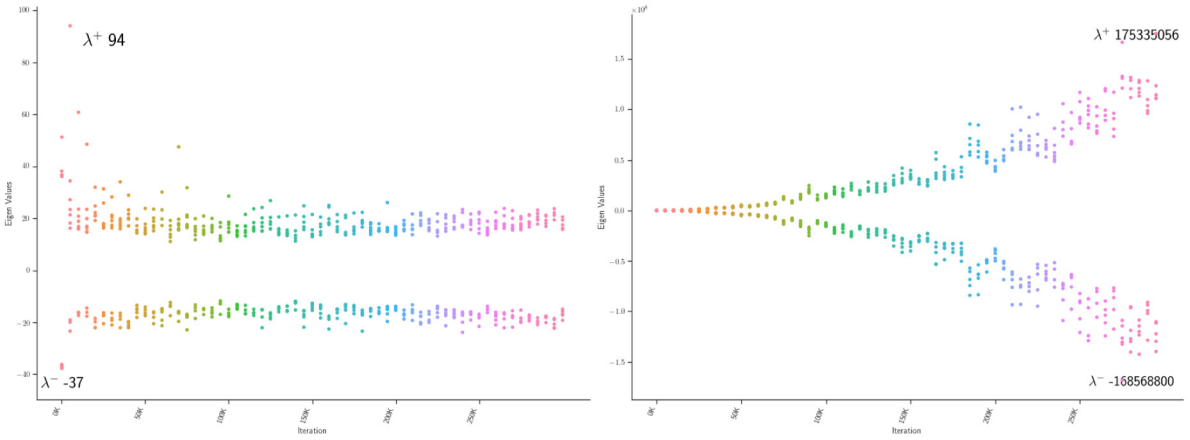
3.2.3. Explaining the dual video discriminator

Stochastic gradient decent (SGD) requires a smooth loss landscape for stable optimization. An ill-conditioned Hessian alludes to directions of high curvature in this landscape that lead to instabilities in the optimization of typical neural networks (Martens, 2010; Saarinen, Bramley, & Cybenko, 1993). The GAN optimization process is itself also highly unstable, exhibiting rotational mechanics and cyclical dynamics that are detrimental to convergence (Balduzzi et al., 2018), more so if the true data distribution is concentrated on a lower dimensional manifold as is likely in video (Mescheder, Geiger, & Nowozin, 2018; Nagarajan & Kolter, 2017).

Fig. 2 shows that the optimization landscape induced by video GAN discriminators has significant pathologies. For the MoCoGAN discriminators, the highest Hessian eigenvalue observed is at times up to an order of magnitude larger than the next leading eigenvalue. This behaviour is observed at the early stages of training for the 2D image discriminator but these extreme outliers exist throughout training for the 3D video discriminator. More importantly, Fig. 2 shows that these pathologies are made worse as the kernel dimensionality of the discriminator increases. The eigenvalues of the loss Hessian induced by the video discriminator are altogether almost an order of magnitude larger than those of the image discriminator. Fig. 3 shows that this is irrespective of dataset and kernel parameter complexity. Altogether, these observations help to explain the emergence of dual 2D image and 3D video discriminators in models such as MoCoGAN (Tulyakov et al., 2018). SGD and its derivatives face a bigger challenge optimizing the loss landscape induced by a 3D discriminator when compared to that of a 2D discriminator. As a result, the 2D discriminator likely improves performance by providing a better image-level gradient signal for the generator. The observations in Table 2, row 6, where increasing the number of parameters for the 3D video discriminator did not lead to a significant boost in performance lend further support to this hypothesis.

3.3. TGAN discriminator

In the previous section, we established that naive application of 3D discriminators induces loss landscapes with high curvature when compared to that induced by 2D discriminators. TGAN (Saito et al., 2017) utilizes a single 3D video discriminator and manages to match the performance of dual discriminator models like MoCoGAN. In our replication study (see Table 8); TGAN consistently outperforms MoCoGAN. Inspired by Wasserstein GAN (Arjovsky et al., 2017), TGAN proposes clamping of the



(a) TGAN - SVC (b) TGAN - Normal

Fig. 4. The 10 leading eigenvalues of the loss Hessian with respect to the TGAN discriminator, with and without SVC applied. λ^+ denotes the largest positive eigenvalue encountered during training and λ^- the largest negative eigenvalue.

Table 3
TGAN Model.

(a) TGAN Reproduction

Model	IS ↑	FID ↓
TGAN — SVC	11.93 ± .08	9127.80 ± 13.77
TGAN – SVC – Original (Saito et al., 2017)	11.85 ± .07	
TGAN — Normal	8.98 ± .06	10093.02 ± 11.06
TGAN – Normal – Original (Saito et al., 2017)	9.18 ± .11	
TGAN – SVC – 2xTime	13.28 ± .09	8797.95 ± 10.39

(b) TGAN Discriminator

Layer	Block Configuration
Input	16 × height × width × 3
c0	Conv3D-(N64, K4, S2, P1), LReLU
c1	Conv3D-(N128, K4, S2, P1), BN, LReLU
c2	Conv3D-(N256, K4, S2, P1), BN, LReLU
c3	Conv3D-(N512, K4, S2, P1), BN, LReLU
c4	Conv2D-(N1, K4, S1, P0,)

spectral norm of the discriminator to a maximum of one in order to stabilize training. This is achieved in practice via Singular Value Clipping (SVC) of the weight matrices such that all singular values are equal to or less than one, enforcing a 1-Lipschitz constraint on the video discriminator. This clipping is applied every *n* iterations during training and the TGAN authors demonstrate that applying SVC leads to stable training and significantly better performance. We analyse the loss landscape induced by SVC and observe how it impacts model performance. The results are presented in Table 3 and Fig. 4.

Fig. 4 provides an interesting insight into the TGAN discriminator, especially when contrasted against the MoCoGAN video discriminator (see Fig. 2(b)). The TGAN discriminator induces a loss landscape filled with saddle points, characterized by the symmetry between the positive and negative eigenvalues of the loss Hessian. Without SVC, we observe that the curvature of the loss landscape becomes more extreme throughout training (Fig. 4(b)). With SVC applied, we observe a comparatively smooth loss landscape with a better conditioned Hessian. This leads to more stable training dynamics and better performance as shown in Table 3a.

We also observe from results presented in row 2 of Table 2 and the last row of Table 3a, that temporal subsampling of video frames has a significant impact on model performance. When temporal subsampling of frames is experimentally controlled for, it appears that TGAN significantly outperforms MoCoGAN according to both FID and IS.

4. Lower-dimensional video discriminators

We have established that good discriminator performance is promoted by a smooth loss landscape, which is well understood for neural networks. We have shown how enforcing a 1-Lipschitz constraint on the discriminator can smooth the loss landscape. We have also demonstrated that there is a strong correlation between the conditioning number of the Hessian and the kernel dimensionality of the discriminator. This opens up an interesting direction in terms of discriminator architecture design.

- *"Do video discriminators require 3D kernels?"*

In addition to the higher curvature optimization landscapes induced by higher dimensional discriminators, there are other disadvantages to using higher dimensional kernels such as an increase in computation and memory costs. In this section, we propose solutions to these issues by exploiting the insights gained from Section 3.

Most generators for video GANs utilize kernels with a maximum dimensionality of two (Saito et al., 2017; Saito & Saito, 2018; Tulyakov et al., 2018), but all discriminators currently incorporate 3D kernels. Thus, we explore the possibility that it may be possible to capture temporal dynamics using a more compressed kernel representation since locally, most information useful for video discrimination may lie on a lower dimensional manifold. This hypothesis is supported by results from related domains such as video recognition, where (Feichtenhofer, Pinz, & Wildes, 2016; Lin, Gan, & Han, 2018; Qiu, Yao, & Mei, 2017; Sun, Jia, Yeung, & Shi, 2015; Tran et al., 2018; Xie, Sun, Huang, Tu, & Murphy, 2018) have successfully removed 3D kernels from classification models without compromising model performance. In most cases, performance has improved and our observations from Section 3 provide a possible explanation for this phenomenon, a better conditioned loss Hessian. Similarly, we seek to replace 3D video GAN discriminators with lower dimensional approximations, resulting in memory and computational efficiency gains as well as better performance due to more stable training dynamics.

We now introduce a family of Lower Dimensional Video Discriminators for Generative Adversarial Networks (LDVD-GANs).

Table 4
Factorized discriminators.

(a) MoCoGAN — Factorized

Layer	Configuration	Block Operations
$c0_{h,w}$	K(1,4,4),S(1,2,2),P(0,1,1),ch64	$c0_{h,w}$, LReLU
$c0_t$	K(4,1,1),S1,P(1,0,0),ch64	$c0_t$, LReLU
$c1_{h,w}$	K(1,4,4),S(1,2,2),P(0,1,1),ch128	$c1_{h,w}$, LReLU
$c1_t$	K(4,1,1),S1,P(1,0,0),ch128	$c1_t$, BN, LReLU
$c2_{h,w}$	K(1,4,4),S(1,2,2),P(0,1,1),ch256	$c2_{h,w}$, LReLU
$c2_t$	K(4,1,1),S1,P(1,0,0),ch256	$c2_t$, BN, LReLU
$c3_{h,w}$	K(1,4,4),S(1,2,2),P(0,1,1),ch256	$c4_{h,w}$, LReLU
$c3_t$	K(4,1,1),S1,P(1,0,0),ch1	$c4_t$

(b) TGAN — Factorized

Layer	Configuration	Block Operations
$c0_{h,w}$	K(1,4,4),S(1,2,2),P(0,1,1),ch64	$c0_{h,w}$, LReLU
$c0_t$	K(4,1,1),S(2,1,1),P(1,0,0),ch64	$c0_t$, LReLU
$c1_{h,w}$	K(1,4,4),S(1,2,2),P(0,1,1),ch128	$c1_{h,w}$, LReLU
$c1_t$	K(4,1,1),S(2,1,1),P(1,0,0),ch128	$c1_t$, BN, LReLU
$c2_{h,w}$	K(1,4,4),S(1,2,2),P(0,1,1),ch256	$c2_{h,w}$, LReLU
$c2_t$	K(4,1,1),S(2,1,1),P(1,0,0),ch256	$c2_t$, BN, LReLU
$c3_{h,w}$	K(1,4,4),S(1,2,2),P(0,1,1),ch512	$c3_{h,w}$, LReLU
$c3_t$	K(4,1,1),S(2,1,1),P(1,0,0),ch512	$c3_t$, BN, LReLU
$c4_{h,w}$	K(4,4),S1,P0,ch1	$c4_{h,w}$, LReLU

These discriminators are characterized by having a maximal kernel dimension that is lower than the ambient dimension of the data modality they are applied to. The ambient dimension for video data is 3D, thus LDVD-GANs are restricted to 1D and 2D kernels, with 2D kernels being the kernels of maximum dimension.

4.1. Factorized convolutions

Decomposing 3D convolution kernels into 2D and/or 1D is an area of active research interest in video recognition and understanding (Feichtenhofer et al., 2016; Qiu et al., 2017; Sun et al., 2015; Tran et al., 2018; Xie et al., 2018). This process can formally be defined as:

$$\mathbf{K}^{h,w,t} = \mathbf{A}^{h,w} \otimes \mathbf{b}^{t}. \qquad \text{where } \mathbf{K} \in \mathbb{R}^{k_h \times k_w \times k_t}, \mathbf{A} \in \mathbb{R}^{k_h \times k_w}, \mathbf{b} \in \mathbb{R}^{k_t}, \tag{2}$$

are convolution kernels and $\otimes$ denotes the Kronecker product. $\mathbf{K}$ is from the subset of kernels that can be factorized into $\mathbf{A}$ and $\mathbf{b}$ as shown in Eq. (2). A convolution over a feature map $\mathbf{F} \in \mathbb{R}^{f_h \times f_w \times f_t}$ can then be defined as:

$$\mathbf{F}^{i+1} = (\mathbf{F}^{i}_{f_h, f_w} \circledast \mathbf{A})_{f_t} \circledast \mathbf{b} \tag{3}$$

where $\circledast$ denotes the convolution operation[9] and the subscripts denote the dimensions over which it is applied.

We observe that some work in orthogonal domains such as video classification has explored similar factorizations to improve performance. In order to situate this article in the video classification domain, we now review all the relevant literature in this field.

The R(2+1)D model (Tran et al., 2018) integrates batch normalization and activation layers after every convolution and applies a factorization similar to Eq. (3) sequentially, while the S3D-G (Xie et al., 2018) model uses similar factorizations within its inception block and modulates the information processed through them via a feature gating mechanism. The Pseudo-3D (P3D) family of models (Qiu et al., 2017) explores different orderings of spatial and temporal convolutions within the bottleneck block of a 2D residual network. The $F_{ST}CN$ model (Sun et al., 2015) splits the network in half, applying spatial convolutions to the first half and temporal convolutions to the down-stream features. In the two-stream literature, ST-ResNet (Feichtenhofer et al., 2016) applies a temporal-spatial–temporal factorization within its bottleneck blocks combined with inter-stream residual paths. All of these models present impressive results when compared to their 3D counterparts within the domain of video classification.

The factorized convolution applied in our discriminators is unique but can be related to the one shown in Eq. (3). Our factorized convolutions are applied via 3D convolution layers. For spatial convolutions, the temporal filter size is set to one. Its output is passed through an activation layer and the resulting feature map is passed through another 3D kernel with its spatial filter size set to 1. This filter aggregates and processes temporal information. It is important to note that unlike in other fields, the intermediate activation layer between the spatial and temporal convolution operations is crucial for video GAN performance and we do not apply batch normalization before or after it.

The factorized MoCoGAN and TGAN discriminator architectures are presented in Table 4. They are identical to their original counterparts with the exception that all the 3D convolution kernels are factorized.

4.2. Temporal shift module

The Temporal Shift Module (TSM) (Lin et al., 2018), entirely forgoes 3D and/or 1D kernels. Instead, it adapts a purely 2D network for video processing by shifting a portion of channels temporally. This allows for an increase in the temporal receptive field of each layer controlled by the temporal shift distance and the layer depth.

For MoCoGAN, we apply the TSM to the MoCoGAN image discriminator architecture and use that as the sole discriminative function during training. For TGAN, we first replace all 3D convolution layers with their 2D counterparts and then interleave convolution operations with shifting operations via the use of TSMs.

Temporal shifting is only applied between intermediate layers and the temporal shift distance in either direction is a single time-step applied to a quarter of the channels for each direction as in Lin et al. (2018). We do not integrate any residual or skip connections before, during or after shifting operations.

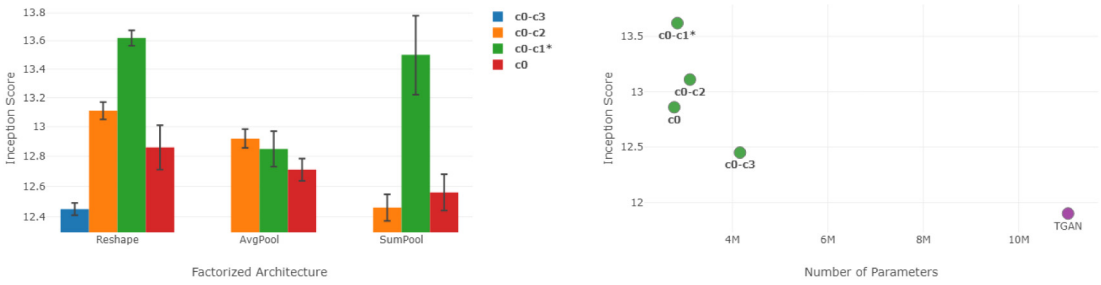
5. Experiments

It is important to note that we do not do any hyper-parameter tuning, nor do we deviate from the experimental setups of the original TGAN and MoCoGAN experiments. Our sole focus in these experiments is to apply the insights gained from Section 3 and demonstrate the efficacy of using Lower Dimensional Video Discriminators (LDVDs) for video generation. In doing so we also improve on the state-of-the-art for video generation and provide a significantly more efficient architecture for high-resolution video generation, competitive with state-of-the-art multi-gpu models. Crucially, we achieve all this by showing that our proposed lower dimensional discriminators can double the performance of previously published models and set state-of-the-art results using only a single GPU.

5.1. Factorized convolutions

We explore the space of lower dimensional video architectures induced by our formulation in Section 4. We aim to gauge

[9] This includes all operations associated with deep learning convolutions; e.g. padding, dilation, etc.



(a) Pooling strategy vs Performance

(b) Parameter Efficiency vs Performance

Fig. 5. Performance of factorized TGAN discriminator architectures on 64 × 64 UCF-101 video generation.

Table 5
Performance on the UCF-101 dataset for discriminators with factorization applied to a varying number of convolution layers.

(a) TGAN — Factorized Convolutions

Model	Inception Score↑	FID↓	Params
TGAN (Saito et al., 2017)	11.85 ± .07		11M
TGAN Ours	11.93 ± .08	9128 ± 14	11M
Layer: c0	**12.86 ± .15**	**9031 ± 2**	2.8M
Layer: c0–c1	**13.62 ± .06**	**8943 ± 4**	2.8M
Layer: c0–c2	**13.11 ± .06**	**8989 ± 4**	3.1M
Layer: c0–c3	**12.45 ± .04**	**9082 ± 1**	4.2M

(b) MoCoGAN — Factorized Convolutions

Model	Inception Score↑	FID↓	Params
MoCoGAN (Tulyakov et al., 2018)	12.42 ± .03		3.3M
MoCoGAN Ours	11.58 ± .04	9485 ± 15	3.3M
Layer: c0	9.71 ± .05	9886 ± 13	0.7M
Layer: c0–c1	**11.60 ± .05**	**9424 ± 12**	0.7M
Layer: c0–c2	10.56 ± .10	9518 ± 6	1M
Layer: c0–c3	10.20 ± .09	9694 ± 3	1M

how much dimension factorization affects baseline performance when using a lower dimensional video discriminator. As such, we benchmark lower dimensional discriminators that are factorized to different degrees. Our proposed LDVD architectures in Table 4 apply factorized convolutions for all 3D convolution layers in their respective video discriminator architectures. For the MoCoGAN discriminator (Table 1), factorizing layer c0 to c3 corresponds to the factorized discriminator architecture shown in Table 4a whose performance is presented in the final row of Table 5b. The same applies for the TGAN discriminator (Table 3b), where factorizing layer c0 to c3 corresponds to the factorized discriminator architecture shown in Table 4b whose performance is presented in the final row of Table 5a. We denote the different discriminators by the layers at which factorization is applied. c0 denotes factorization of the first 3D convolution layer, c0–c1 denotes factorization of the first and second convolution layers, so on and so forth. All other non-factorized convolution layers are restricted to using 2D convolution kernels. The performance of models trained with these discriminators is presented in Table 5. The TGAN discriminator has an additional 2D layer appended to it. We explore how temporal aggregation methods applied to its inputs affect model performance. The results are presented in Fig. 5.

Table 5 and Fig. 5 provide for several interesting observations; the first being that factorized LDVDs can outperform their 3D counterparts using a fraction of their parameters. In particular for the TGAN discriminator, performance improves in every case. The best performing factorized TGAN discriminator boosts the IS by around 15% and consequently the state-of-the-art for 64 × 64 video generation by around 10%. In comparison, we only observe moderate performance improvements for the MoCoGAN model and suspect that this is likely due to a bottlenecked discriminator. We denote the best performing factorized TGAN and MoCoGAN models as TGAN-F and MoCoGAN-F respectively. These are the c0–c1 discriminators from Table 5.

Improving MoCoGAN performance. In Section 3 we demonstrate three ways of encouraging smooth loss landscapes; increasing the number of discriminator parameters, using an LDVD and directly enforcing that all layers in the discriminator are 1-Lipschitz continuous. The TGAN model applies Singular Value Clipping (SVC) to stabilize training and its factorized discriminators benefit from it. We explore enforcing a 1-Lipschitz constraint on the MoCoGAN-F discriminator via spectral normalization (Miyato et al., 2018) and observe that it improves performance to an IS of 12.33 ±.09 and an FID of 9069 ± 4.

Temporal resolution. An interesting observation about the MoCoGAN-F and TGAN-F discriminators is that they only process temporal information in their initial couple of layers. Factorized discriminators with more capacity to process temporal information further downstream and over a longer temporal receptive field do not perform as well.

5.2. Temporal shift module

Table 6 summarizes experiments from our exploration of the temporal shifting strategy. As in the original TSM models from Lin et al. (2018), single-step temporal shifting in each direction is applied to a quarter of feature maps. Our discriminator naming convention is similar to that in the previous section, whereby ci–ck denotes a discriminator with temporal shifting applied after layers i through k. All discriminators are entirely 2D in nature with temporal shifting used to capture and merge temporal information between intermediate layers.

The temporal shift module (TSM) allows for video discriminators with the same number of parameters as traditional image discriminators. Additionally, the parameter cost of the discriminator is constant regardless of the size of the temporal receptive field. This can be seen in Table 6a, where a 2D discriminator improves the performance of the TGAN model by around 10%, through depth-wise regulation of the Temporal Receptive Field (TRV). We observe a drop in performance for the MoCoGAN model (Table 6b) and attribute it to the discriminator not being 1-Lipschitz continuous. We denote the best performing discriminators for TGAN and MoCoGAN from Table 6, TGAN-TSM and MoCoGAN-TSM respectively.

When comparing temporal shifting to factorized convolutions, we observe that the temporal processing carried out by the 1D

Table 6
Performance on the UCF-101 dataset for discriminators using the Temporal Shift Modules at different layer depths.

(a) TGAN — Temporal Shifting

Model	Inception Score↑	TRV	Params
TGAN (Saito et al., 2017)	11.85 ± .07	22	11M
TGAN Ours	11.93 ± .08	22	11M
Layer: c0	11.76 ± .06	3	2.8M
Layer: c0–c1	**12.11 ± .12**	5	2.8M
Layer: c0–c2	**12.32 ± .07**	7	2.8M

(b) MoCoGAN — Temporal Shifting

Model	Inception Score↑	TRV	Params
MoCoGAN (Tulyakov et al., 2018)	12.42 ± .03	22	3.3M
MoCoGAN Ours	11.58 ± .04	22	3.3M
Layer: c0	8.90 ± .05	3	0.7M
Layer: c0–c1	9.82 ± .07	5	0.7M
Layer: c0–c2	9.60 ± .02	7	0.7M

kernels in factorized convolutions result in significant improvements in performance for negligible cost in terms of memory and computation. The shifting strategy is comparatively expensive, often increasing training times by 10%–40% depending on the number of shifting operations carried out.

5.3. Comparison with state-of-the-art

5.3.1. A note on the IS and FID of UCF-101 videos

An accurate comparison against previous work requires that a distinction be made between low and high resolution video generation. This is because metrics such as the Inception Score (IS) and Fréchet Inception Distance (FID) are known to be sensitive to image resolution (Borji, 2019; Brock et al., 2019). In video, the additional temporal dimension is unconstrained and requires some procedure for sub-sequence selection. Additionally, data augmentation methods such as frame sub-subsampling (i.e. skip every *n* frames) or cropping are applied to some published models but not others. We benchmark the IS of the UCF-101 *'trainlist01'* dataset used in the video GAN literature under different conditions. Since dataset videos have more frames than the evaluation networks 16 frame temporal resolution, we can derive a rough standard deviation by repeating the evaluation process four times with randomly sampled 16 frame sub-sequences. These results are shown in Table 7.

Dataset normalization. We observed that both the IS and FID are highly sensitive to the mean normalization file used during evaluation and as a result maintain the use of the mean normalization file provided by the TGANv2 authors (Saito & Saito, 2018). We note that our 128 resolution results in Table 7 are close to the true data IS of 83.18 published in Acharya et al. (2018).

5.3.2. Lower resolution generation

Table 8 presents results for the best performing TSM and factorized discriminators against the state-of-the-art models for 64 × 64 video generation.

Both TGAN-TSM and TGAN-F outperform the original TGAN architecture using a quarter of the parameters of the original discriminator. Furthermore, TGAN-F sets a new state-of-the-art result for 64 × 64 video generation and outperforms complex higher resolution models such as ProVGAN (see Table 9) without exploiting temporal subsampling, a technique which we have shown to significantly boost performance (see the second row of Table 2 and the last row of Table 3a). Next, we explore higher resolution video generation with this architecture.

5.3.3. Higher resolution generation

The computation and memory gains achieved by reducing the maximum kernel dimension of the discriminator allows for the TGAN model to be scaled up to higher resolutions without issue, even on a single GPU system. Our high-resolution video generation model is based on our best performing low-resolution model, TGAN-F, with appropriate modifications made to support higher resolutions.

Table 9 presents results for TGAN-F benchmarked on the task of 128 × 128 video generation. TGAN-F (+ 4xTemporalCh) corresponds to quadrupling the number of channels in the temporal frame generator, TGAN-F (+ 2xImageCh) corresponds to doubling the number of channels in the TGAN-F image generator, TGAN-F (+ 2xTime) corresponds to doubling the temporal receptive field by subsampling a longer video and TGAN-F (+ All) corresponds to applying all the above modifications to the TGAN-F architecture.

The first observation is that TGAN-F trained on a single GPU outperforms all single-GPU models by at least 15%. The second observation is that TGAN-F is significantly more parameter efficient than previous video GAN architectures. This consequently enables it to be more memory efficient, enabling training with batch sizes of up to 32 and at resolutions as high as 16 × 128 × 128 on a single GPU. Another observation is that TGAN-F (+ All) almost doubles the performance of the original TGAN model while using the same generator architecture, hyper-parameters and hardware constraints. The last observation is that TGAN-F (+ All) on a single gpu provides for state-of-the-art results while using a fraction of the parameters and compute of multi-GPU VGAN models. This demonstrates the efficacy and efficiency of LDVD-GANs like TGAN-F, and shows its superior performance when compared to the original TGAN model and many other video GAN models. TGANv2 (Saito & Saito, 2018), is the state-of-the-art model for video generation in large distributed multi-GPU settings, where it achieves an IS of 54.93. Under similar computational restraints, its performance is still worse than an older model trained with an LDVD, this demonstrates the efficacy of our proposed discriminator design. It should be noted that TGANv2 is unable to match our single-GPU performance even when the model has access to more than two times the GPU memory available to TGAN-F. Additionally TGANv2 generates video at a higher resolution, and the IS metric is biased in favour of higher resolutions (see Table 7 for the positive effect of resolution on the IS metric). TGANv2 trained under the exact same experimental conditions as our proposed models (see row 6 of Table 9), can only fit a batch-size of 2 videos on the GPU at any one time and results in an IS of 14.04. Our most basic LDVD model achieves an IS of 16.85 while using less than a tenth of the parameters and can train with batch sizes 16 times larger under the same computational constraints. Our best model achieves an Inception Score of 22.91, a 1.5x improvement in performance over TGANv2, with less than half the parameters and 16 times the batch-size on a single GPU.

TGAN-F loss landscape. Hessian analysis of the loss landscape induced by the TGAN-F discriminator during optimization shows that the curvature of this space is more than halved when compared to that of the original TGAN discriminator (see Figs. 6(b) vs 6(a)). The smoother loss landscape induced by the lower dimensional discriminator helps to explain the improved performance of the TGAN-F model when compared to its original higher dimensional counterpart, TGAN.

5.3.4. Qualitative results

Fig. 8 shows frames from TGAN-F models trained on MUG-FED at different resolutions. Video generation samples at resolutions

Table 7
Inception Score and FID of the UCF-101 dataset under different conditions.

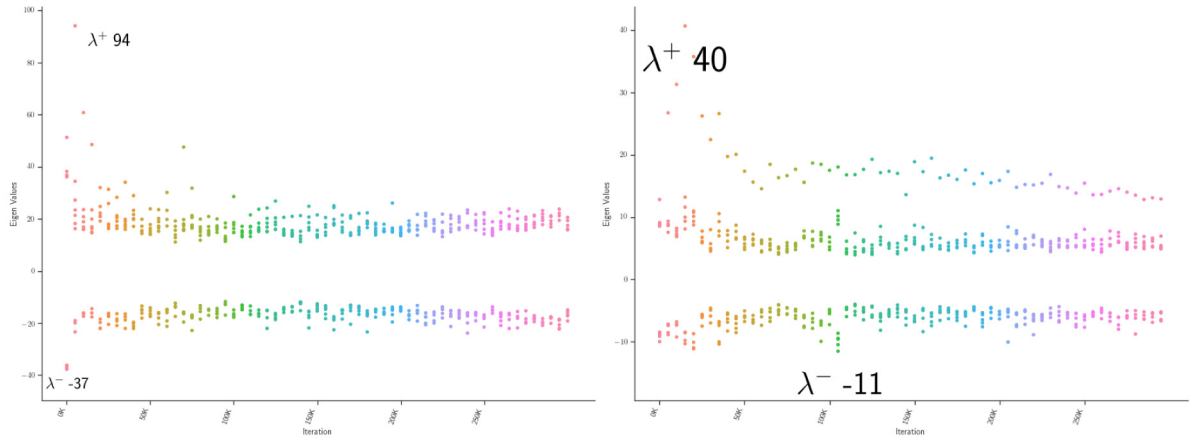
Inception Score	FID	Resolution	2x Subsampling	Random Video Reversal	Crop
70.01 ± .08	1765.86 ± 2.66	16 × 64 × 64 × 3			Centre
73.63 ± .11	1685.63 ± 2.73	16 × 64 × 64 × 3	✓		Centre
73.46 ± .17	1686.85 ± 1.71	16 × 64 × 64 × 3	✓	✓	Centre
71.68 ± .21	1652.62 ± 1.07	16 × 64 × 64 × 3	✓	✓	Random
94.49 ± .06	728.63 ± 1.79	16 × 128 × 128 × 3			Centre
94.93 ± .09	804.19 ± 0.91	16 × 128 × 128 × 3	✓		Centre
94.66 ± .11	804.40 ± 0.67	16 × 128 × 128 × 3	✓	✓	Centre
94.16 ± .11	718.91 ± 1.54	16 × 128 × 128 × 3	✓	✓	Random

Table 8
Performance on the 64 × 64 UCF-101 video generation benchmark.

Model	Inception Score ↑	FID ↓	Parameter Reduction ↑
VGAN (Vondrick et al., 2016) (2016)	8.18 ± .05		
TGAN (Saito et al., 2017) (2017)	11.85 ± .07		
MoCoGAN (Tulyakov et al., 2018) (2018)	12.42 ± .03		
TGAN (our reproduction)	11.93 ± .08	9127.80 ± 13.77	
TGAN-TSM	12.32 ± .07	9796.68 ± 2.29	74.93%
TGAN-F	**13.62 ± .06**	**8942.63 ± 3.72**	74.19%
MoCoGAN (our reproduction)	11.58 ± .04	9485.34 ± 14.61	
MoCoGAN-TSM	9.82 ± .07	10608.66 ± 15.67	79.99%
MoCoGAN-F	11.60 ± .05	9424.01 ± 12.16	69.61%
MoCoGAN-F + SN	12.33 ± .09	9069.11 ± 3.97	69.61%
MoCoGAN + TGAN-F Discriminator	**12.53 ± .01**	**9038.42 ± 9.21**	15.16%

Table 9
Performance on the UCF-101 benchmark for high-resolution video generation.

Model	Resolution	Batch size	GPU/Memory	IS↑	Params
ProVGAN (Acharya et al., 2018) (2018)	32 × 256 × 256		Multi-GPU/32 GB	13.59	
ProVGAN + SWGAN (Acharya et al., 2018) (2018)	32 × 256 × 256		Multi-GPU/32 GB	14.56	
TGANv2 + 2xTime (Saito & Saito, 2018) (2020)	16 × 192 × 192	8	1 GPU/32 GB	20.61 ± .28	200M
TGANv2 + 4xTime (Saito & Saito, 2018) (2020)	16 × 192 × 192	8	1 GPU/32 GB	21.45 ± .29	200M
TGANv2 + 1xTime - (Our run)	16 × 192 × 192	2	1 GPU/12 GB	14.04 ± .34	200M
TGAN-F	16 × 128 × 128	32	1 GPU/12 GB	16.85 ± .04	16M
TGAN-F + 4xTemporalCh	16 × 128 × 128	32	1 GPU/12 GB	17.72 ± .20	27M
TGAN-F + 2xImageCh	16 × 128 × 128	32	1 GPU/12 GB	20.35 ± .23	25M
TGAN-F + 2xTime	16 × 128 × 128	32	1 GPU/12 GB	17.23 ± .15	16M
TGAN-F + All	16 × 128 × 128	32	1 GPU/12 GB	**22.91 ± .19**	70M



(a) TGAN (b) TGAN - F

Fig. 6. The 10 leading eigenvalues of the Hessian with respect to the TGAN and TGAN-F discriminator. λ^+ denotes the largest positive eigenvalue encountered during training and λ^- the largest negative eigenvalue.

of 512 × 512 are available in the supplementary material.[10] Fig. 9 shows high-resolution samples from the same model trained on the UCF-101 dataset. We observe learned camera zooming and panning motions. Full resolution random samples and other qualitative results are available in the supplementary material.

Visual quality. Figs. 8, 9 and 7(d) present the current state-of-the-art visual results. As can be seen from these samples, there is

[10] Supplementary Material: https://drive.google.com/drive/folders/1J9gjS2HRTwoADQVqVoMbs5pBtO7rOGF6.

(a) MoCoGAN (b) MoCoGAN - F with SN

(c) TGAN (d) TGAN - F

Fig. 7. The first frame of video samples for each of the models on 64 × 64 UCF-101. Full samples available in the supplementary material.

Table 10
Human evaluation of the quality and diversity of samples generated by different models trained on the MUG-FED dataset.

Model Comparison	Quality (%)	Diversity (%)
MoCoGAN/TGAN	33.0/67.0	50.0/50.0
TGAN/TGAN-F	41.6/58.4	50.0/50.0
MoCoGAN/TGAN-F	25.0/75.0	16.6/83.4

still room for improvement. We observe that video GANs suffer from the same fundamental issue as image GANs; capturing and appropriately modelling high-level global structure. The individual patches of an image or video frame, may look realistic, but the overall image likely does not. This behaviour is observed in the best performing image GANs such as BigGAN (Brock et al., 2019), and is exaggerated in video with the addition of the temporal dimension. Unrealistic video frames undergo unrealistic transformations resulting in video that does not look realistic. Possible ways to tackle this issue include encouraging global consistency via pixel level supervision (e.g. a reconstruction loss), or integrating structural priors such as attention (Zhang et al., 2019), or training with augmented video data (e.g. integrating optical-flow).

5.3.5. Quantitative results

A human evaluation was carried out to compare the quality and diversity of samples generated by models using factorized discriminators on the MUG-FED dataset against other models in the literature. These results are presented in Table 10.

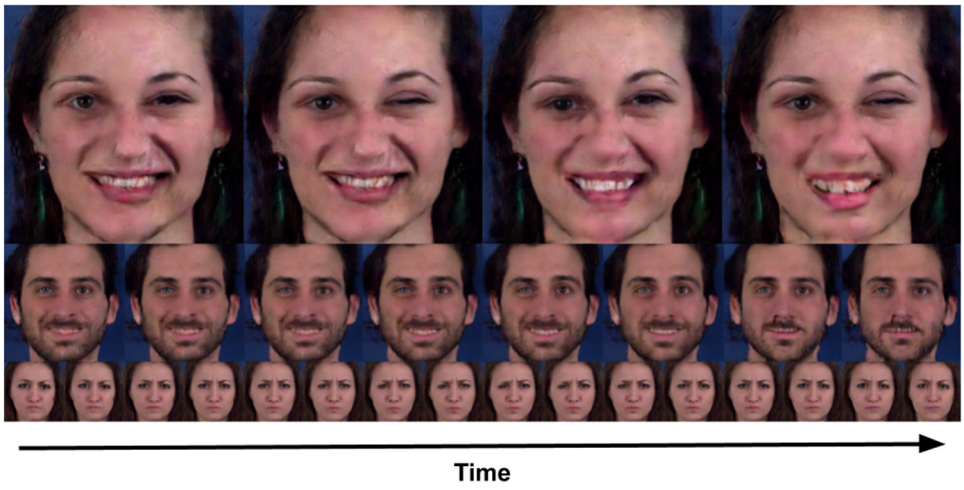


Fig. 8. Frames from videos generated by our TGAN-F model trained on MUG-FED at 256 × 256, 128 × 128, and 64 × 64 resolutions (top to bottom).

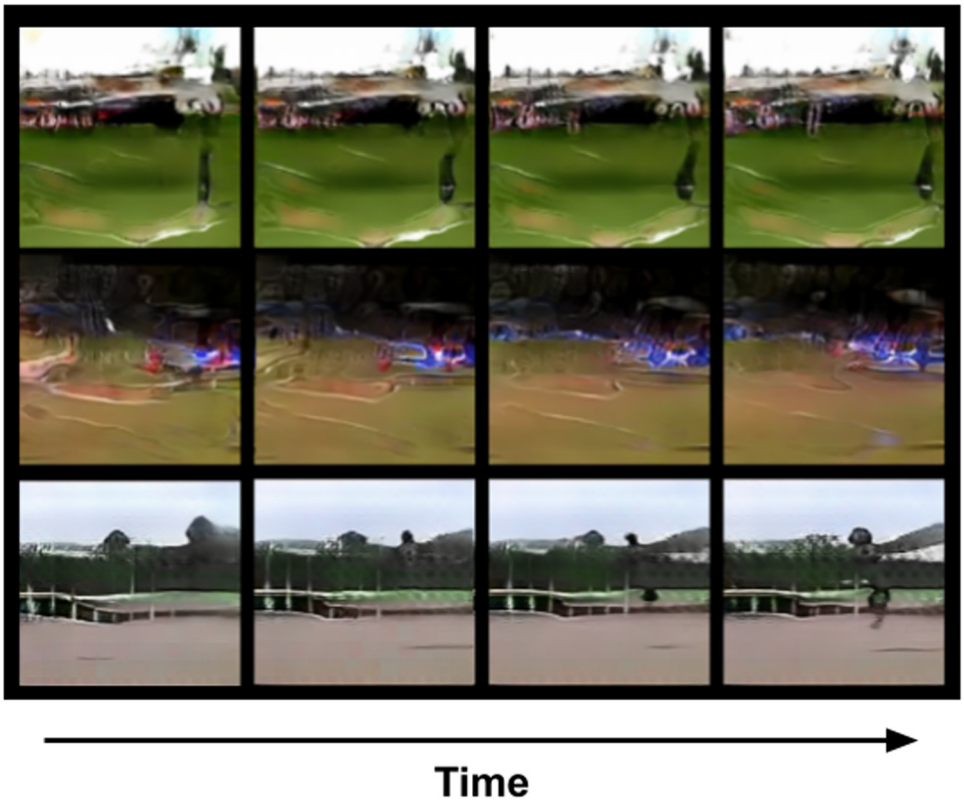


Fig. 9. Frames from videos generated by TGAN-F trained on UCF-101 at 128 × 128 resolutions. Top: Zoom into scene, Middle: Pan left to right, Bottom: Rotate around centre of focus.

6. Conclusion

The field of image generation has enjoyed significant advances in recent years, and our work aims at taking a step towards doing the same for video generation. Specifically, we study the properties of video discriminator architectures and find that higher dimensional video discriminators induce a loss landscape with relatively higher curvature. As a result, we question the utility of 3D kernels in video GAN models and empirically demonstrate that they are not required for the video generation problem as it is currently framed. Our design proceeds by replacing 3D kernels with lower dimensional approximations, and our proposed lower dimensional discriminators, improve the performance of video GAN generators they are applied to. As a result, we demonstrate performance that is competitive with the state-of-the-art for both single and multi-gpu video generation; in both low-resolution and high-resolution video generation settings.

We carried out a wide range of experiments across two generator models and many more discriminator architectures. We summarize the successful results of this investigation in Sections 3 and 5. These experiments demonstrate that the curvature of the loss landscape for video GAN discriminators increases

with kernel dimensionality. We also uncover guiding principles to limit this behaviour; mainly not only avoiding 3D kernels, but also enforcing a 1-Lipschitz discriminator and increasing the number of parameters in a model (Section 4). Based on these principles, we propose a family of lower dimensional video discriminator architectures that provide for efficient but powerful video GAN models. Subsequently, we explore one such lower dimensional discriminator architecture, TGAN-F, resulting in state-of-the-art performance for a single-gpu model (Section 5.3).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Acharya, D., Huang, Z., Paudel, D. P., & Gool, L. V. (2018). Towards high resolution video generation with progressive growing of sliced Wasserstein GANs. CoRR abs/1810.02419, URL arXiv:1810.02419.

Aifanti, N., Papachristou, C., & Delopoulos, A. (2010). The MUG facial expression database. In *11th international workshop on image analysis for multimedia interactive services* (pp. 1–4). URL http://ieeexplore.ieee.org/document/5617662/.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th international conference on machine learning* (pp. 214–223). URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., & Graepel, T. (2018). The mechanics of n-player differentiable games. In J. Dy, & A. Krause (Eds.), *Proceedings of machine learning research*: *vol. 80*, *Proceedings of the 35th international conference on machine learning* (pp. 354–363). Stockholmsmässan, Stockholm Sweden: PMLR, URL http://proceedings.mlr.press/v80/balduzzi18a.html.

Barratt, S., & Sharma, R. (2018). A note on the inception score. In *In ICML workshop on theoretical foundations and applications of deep generative models*. URL arXiv:1801.01973.

Borji, A. (2019). Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, *179*, 41–65. http://dx.doi.org/10.1016/j.cviu.2018.10.009, URL http://www.sciencedirect.com/science/article/pii/S1077314218304272.

Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *International conference on learning representations*. URL https://openreview.net/forum?id=B1xsqj09Fm, arXiv:1809.11096.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1179, URL https://www.aclweb.org/anthology/D14-1179.

Feichtenhofer, C., Pinz, A., & Wildes, R. P. (2016). Spatiotemporal residual networks for video action recognition. In *Advances in neural information processing systems 29: Annual conference on neural information processing systems 2016* (pp. 3468–3476). URL http://papers.nips.cc/paper/6433-spatiotemporal-residual-networks-for-video-action-recognition.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., & Bengio, Y. (2014). Generative adversarial networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems (vol. 27)* (pp. 2672–2680). Curran Associates, Inc., URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of wasserstein GANs. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 5769–5779). USA: Curran Associates Inc., URL http://dl.acm.org/citation.cfm?id=3295222.3295327.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017* (pp. 6629–6640). URL https://arxiv.org/abs/1706.08500.

Jolicoeur-Martineau, A. (2019). The relativistic discriminator: a key element missing from standard GAN. In *International conference on learning representations*. URL https://openreview.net/forum?id=S1erHoR5t7.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *International conference on learning representations*. URL https://openreview.net/forum?id=Hk99zCeAb.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *2019 IEEE conference on computer vision and pattern recognition*. URL https://arxiv.org/abs/1812.04948.

Lin, J., Gan, C., & Han, S. (2018). Temporal shift module for efficient video understanding. CoRR abs/1811.08383, URL arXiv:1811.08383.

Martens, J. (2010). Deep learning via Hessian-free optimization. In *Proceedings of the 27th international conference on international conference on machine learning* (pp. 735–742). USA: Omnipress, URL http://dl.acm.org/citation.cfm?id=3104322.3104416.

Mescheder, L., Geiger, A., & Nowozin, S. (2018). Which training methods for GANs do actually converge? In J. Dy, & A. Krause (Eds.), *Proceedings of machine learning research*: *vol. 80*, *Proceedings of the 35th international conference on machine learning* (pp. 3481–3490). Stockholmsmässan, Stockholm Sweden: PMLR, URL http://proceedings.mlr.press/v80/mescheder18a.html.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, conference track proceedings*. URL https://openreview.net/forum?id=B1QRgziT-.

Nagarajan, V., & Kolter, J. Z. (2017). Gradient descent GAN optimization is locally stable. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems (vol. 30)* (pp. 5585–5595). Curran Associates, Inc., URL http://papers.nips.cc/paper/7142-gradient-descent-gan-optimization-is-locally-stable.pdf.

v. Neumann, J. (1928). Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, *100*(1), 295–320, URL https://doi.org/10.1007/BF01448847.

Nowozin, S., Cseke, B., & Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems (vol. 29)* (pp. 271–279). Curran Associates, Inc., URL http://papers.nips.cc/paper/6066-f-gan-training-generative-neural-samplers-using-variational-divergence-minimization.pdf.

Pearlmutter, B. A. (1994). Fast exact multiplication by the Hessian. *Neural Comput.*, *6*(1), 147–160, URL http://dx.doi.org/10.1162/neco.1994.6.1.147.

Qiu, Z., Yao, T., & Mei, T. (2017). Learning spatio-temporal representation with pseudo-3D residual networks. In *IEEE International conference on computer vision* (pp. 5534–5542). URL https://doi.org/10.1109/ICCV.2017.590.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, conference track proceedings*. URL http://arxiv.org/abs/1511.06434.

Saarinen, S., Bramley, R., & Cybenko, G. (1993). Ill-conditioning in neural network training problems. *SIAM Journal on Scientific Computing*, *14*(3), 693–714.

Saito, M., Matsumoto, E., & Saito, S. (2017). Temporal generative adversarial nets with singular value clipping. In *IEEE international conference on computer vision, ICCV 2017, Venice, Italy, October 22-29, 2017* (pp. 2849–2858). URL https://doi.org/10.1109/ICCV.2017.308.

Saito, M., & Saito, S. (2018). TGANv2: Efficient training of large models for video generation with multiple subsampling layers. CoRR abs/1811.09245, URL arXiv:1811.09245.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. In *Advances in neural information processing systems 29: Annual conference on neural information processing systems 2016* (pp. 2226–2234). URL http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.

Schmidhuber, J. (1990). *Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments*: *Technical report*.

Schmidhuber, J. (1991). *Learning factorial codes by predictability minimization*: *Technical report CU-CS-565-91*, Dept. of Comp. Sci., University of Colorado at Boulder.

Schmidhuber, J. (2020). Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991). *Neural Networks*, *127*, 58–66. http://dx.doi.org/10.1016/j.neunet.2020.04.008, URL http://www.sciencedirect.com/science/article/pii/S0893608020301283.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR abs/1212.0402, URL arXiv:1212.0402.

Sun, L., Jia, K., Yeung, D., & Shi, B. E. (2015). Human action recognition using factorized spatio-temporal convolutional networks. In *2015 IEEE international conference on computer vision, ICCV 2015, Santiago, Chile, December 7-13, 2015* (pp. 4597–4605). URL https://doi.org/10.1109/ICCV.2015.522.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018* (pp. 6450–6459). http://dx.doi.org/10.1109/CVPR.2018.00675, URL http://openaccess.thecvf.com/content_cvpr_2018/html/Tran_A_Closer_Look_CVPR_2018_paper.html.

Tulyakov, S., Liu, M., Yang, X., & Kautz, J. (2018). MoCoGAN: Decomposing motion and content for video generation. In *2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018* (pp. 1526–1535). http://dx.doi.org/10.1109/CVPR.2018.00165, URL http://openaccess.thecvf.com/content_cvpr_2018/html/Tulyakov_MoCoGAN_Decomposing_Motion_CVPR_2018_paper.html.

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. In *Advances in neural information processing systems 29: Annual conference on neural information processing systems 2016, December 5-10, 2016, Barcelona, Spain* (pp. 613–621). URL http://papers.nips.cc/paper/6194-generating-videos-with-scene-dynamics.

Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Computer vision - ECCV 2018 - 15th European conference, Munich, Germany, September 8-14, 2018, proceedings, Part XV* (pp. 318–335). http://dx.doi.org/10.1007/978-3-030-01267-0_19, https://doi.org/10.1007/978-3-030-01267-0_19.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In K. Chaudhuri, & R. Salakhutdino (Eds.), *Proceedings of machine learning research*: *vol. 97*, *Proceedings of the 36th international conference on machine learning*. Long Beach, California USA: PMLR,.